Leveraging Fine-Grained Information and Noise Decoupling for Remote Sensing Change Detection

Qiangang Du*
Fudan University
China

Jinlong Peng*
Tencent Youtu Lab
China

Changan Wang
Tencent Youtu Lab
China

Xu Chen
Tencent Youtu Lab
China

Qingdong He
Tencent Youtu Lab
China

Wenbing Zhu
Fudan University
China

Mingmin Chi
Fudan University
China

Yabiao Wang
Tencent Youtu Lab
China

Chengjie Wang
Tencent Youtu Lab
China

ABSTRACT

Change detection aims to identify remote sense object changes by analyzing data between bitemporal image pairs. Due to the large temporal and spatial span of data collection in change detection image pairs, there are often a significant amount of task-specific and task-agnostic noise. Previous effort has focused excessively on denoising, with this goes a great deal of loss of fine-grained information. In this paper, we revisit the importance of fine-grained features in change detection and propose a series of operations for fine-grained information compensation and noise decoupling (FINO). First, the context is utilized to compensate for the fine-grained information in the feature space. Next, a shape-aware and a brightness-aware module are designed to improve the capacity for representation learning. The shape-aware module guides the backbone for more precise shape estimation, guiding the backbone network in extracting object shape features. The brightness-aware module learns a overall brightness estimation to improve the model's robustness to task-agnostic noise. Finally, a task-specific noise decoupling structure is designed as a way to improve the model's ability to separate noise interference from feature similarity. With these training schemes, our proposed method achieves new state-of-the-art (SOTA) results in multiple change detection benchmarks. The code will be made available.

KEYWORDS

Change detection, Fine-grained information compensation, Context-dependent learning

1 INTRODUCTION

Change Detection (CD) is a crucial branch of Earth Remote Sensing (ERS) image analysis, which aims at identifying land surface object changes between bitemporal image pairs. It has widespread applications in detecting changes in ground buildings, land, forests and plays a key role in environmental protection, urban development. These image pairs are acquired by remote sensors in space or in the air at high altitudes over different time periods. Each pixel in the image represents a geographical area and is assigned a binary label indicating whether or not the object in that area has changed. Due to the long time span of image data acquisition, the differences between these bitemporal image pairs often contain a significant amount of noise. These noises can be categorized into task-specific and task-agnostic noise. Task-agnostic noise manifests as overall changes between image pairs, such as changes in seasonality and lighting that cause similar objects to appear differently between the bitemporal image pairs. It presents as intra-class differences. Task-specific noise, on the other hand, exhibits specific semantic changes within regions. For example, objects of different categories exhibit highly similar semantic feature descriptors, the semantic changes between different object classes are not changes of interest. These noises are often highly coupled to the relevant change in the latent space and makes it challenging to separate them.

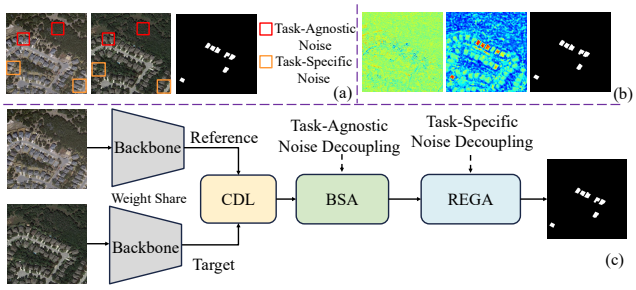

Figure 1: Conceptual illustration of proposed approach. (a) Two types of change detection pseudo-change. (b) Multi-scale fused features (far left) visualised with features learned by FINO (middle), with FINO predictions on the far right. (c) The core process of FINO. Context-dependent learning (DCL) compensates for fine-grained features, brightness-aware and shape-aware (BSA) perception to decouple task-agnostic noise, and regularization gate (REGA) decouples task-specific noise.

Daudt et al. [3] first proposed a Siamese networks based Deep Learning (DL) for change detection, then DL has achieved excellent results in change detection. The current mainstream approach is based on siamese backbones to extract features separately and detect changed objects by the difference information. Recent efforts have focused on designing specialized networks to filter out the mentioned noise. However, by overly focusing on denoising, these

*Equal contribution.

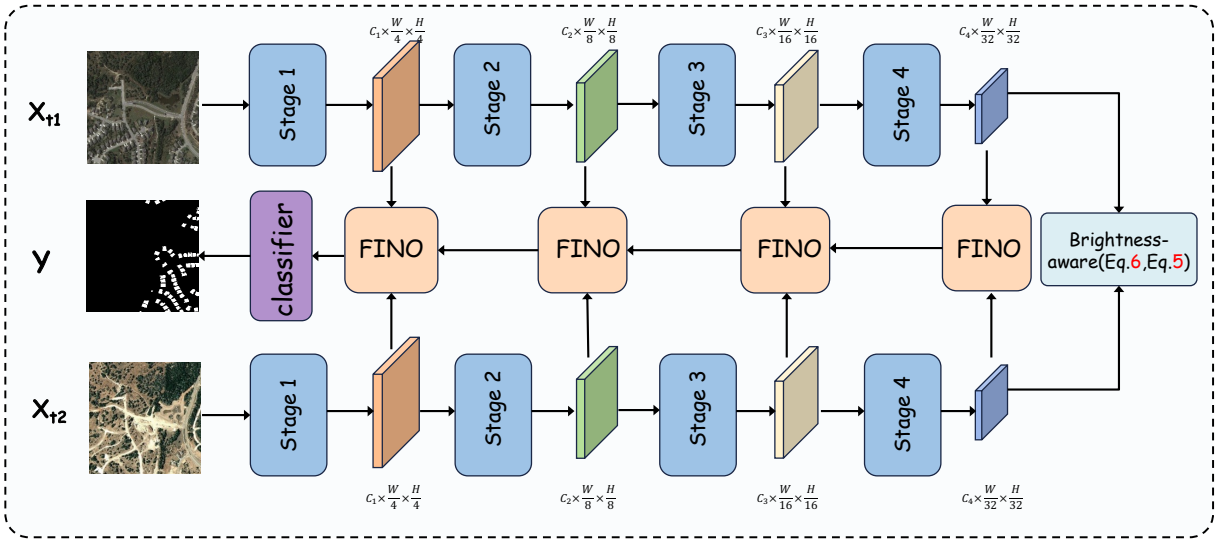

Figure 2: Framework of the proposed network. FINO consists of CDL, shape-aware module, and REGA in tandem. The attention-based CDL adequately compensates for fine-grained information. The shape-aware module decouples the task-agnostic noise and guides the model to learn the object shape representation. REGA decouples the task-specific noise.

methods inevitably suffer from the loss of some crucial detail information, which is essential for change detection, such as location, scale and more. These approaches address this issue by multi-scale features fusion [2, 4–6, 13, 32], feature interaction [4, 15, 27, 30] or skip connections [14, 28] to compensate for fine-grained information. Nevertheless, these methods inevitably lead to the problems of information redundancy and noise regeneration, as shown in Fig. 1(b). Furthermore, as the complexity and depth of the network model increasing, it will lead to overfitting.

In summary of the above review, existing methods overly prioritize denoising and ignore some key issues, as shown in Fig. 1(a). **Firstly**, visual understanding belongs to fine-grained task. Fine-grained details are crucial for identifying the aforementioned pseudo-change noise. However, previous works have largely neglected this aspect and instead solely focused on noise elimination. **Secondly**, global features in change detection primarily manifest in overall descriptions like illumination and seasonality. Simply pursuing long distance feature learning does not significantly enhance detection performance [2, 5]. Change detection exhibits a high degree of local contextual similarity. There is a high probability that an object is surrounded by objects of its own kind. **Finally**, the noise in change detection consists partly of task-specific and task-agnostic components. Treating them as a single entity for denoising while learning features can be disastrous. Extreme data imbalance severely hampers noise suppression, and few works have addressed this issue.

In this paper, we claim that current research in change detection lacks sufficient utilization of fine-grained information and the ability to differentiate noise. We propose a series of operations for **F**ine-grained **I**nformation compensation and **N**oise dec**O**upling (FINO) to improve change detection representation learning, as shown in Fig. 1(c). First, to acquire local features and compensation for fine-grained information, we first employ a context-dependent learning (CDL) module based region attention mechanism to learn contextual semantic features. This approach allows us to gradually reconstruct fine-grained features by using high-level semantic features to guide the reconstruction of lower-level features. Following it, based on alignment operations, we propose a brightness-aware and shape-aware (BSA) module to guide the model in learning object shape descriptions and enhance the model robustness to task-agnostic noise. Finally, we introduce a regularization gate (REGA) structure to decouple task-specific noise. On one hand, the gate structure prevents the noise reintroducted during denoising and improves the confidence of interest object changes. On the other hand, through regularization gating, we deactivate a portion of neurons during gradient computation to enhance the model robustness to task-specific noise.

In summary, the main contributions of this paper are as follows:

- 1. We rethink the importance of fine-grained information in change detection and elucidate the nature of pseudo-change noise. A context-dependent learning structure based regional attention is proposed to compensate for fine-grained information.
- 2. We propose brightness-aware and shape-aware modules to enhance the representation learning of shapes and improve model robustness against task-agnostic noise.
- 3. We introduce a regularization gate structure to decouple task-specific noise which helps mitigate the impact of extreme data imbalance on the model learning in change detection.
- 4. We achieve state-of-the-art performance on multiple change detection benchmarks.

2 RELATED WORK

2.1 Siamese Network for CD

The task of change detection presents unique challenges compared to image classification or segmentation tasks due to its sensitivity to both the bitemporal image features themselves and the difference features that represent whether there are changes. The input to change detection is an image pair consisting of the current temporal image and the past temporal image, making Siamese networks a preferred choice due to their ability to map the respective images into a new space with fewer parameters but the same structure. Most of the existing works [6, 14, 15, 17, 24, 27, 30, 32] use Siamese networks as a backbone. Daudt et al. [3] first proposed Siamese architectures for CD and compared the Early Fusion (EF) and Siamese, proving that last fusion (LF) such as Siamese is superior to EF. However, the inevitable drawback of these methods is that the downsampling causes the loss of detailed information about the data as the depth of the neural network increases. Ziming Li et al. [14, 16] bridge the high-level semantic information and the low-level detailed information through a skip connection structure, thus ensuring that the extracted changed features contain both high-level features and local detailed information. Zhenglai Li et al. [15] use several repetitive multi-scale change information interaction structures to extract change features. Although the above methods enhance the interaction and fusion of information between different branches of Siamese level, they result in information redundancy and regenerate a large amount of noise. It's because that the irrelevant change noises are highly adherent to the changed features on the feature space, these methods that directly adopt raw space features for fusion or interaction would regenerate noise to downstream change detection.

In this paper we refine the change detection noise by decoupling task-specific and task-agnostic noise in REGA and BSA to learn change features more accurately.

2.2 Information Aggregation For CD

Change detection in images often relies on the detailed features of the original image to provide valuable information about the changes. However, existing methods, such as feature difference-based change detection, often fail to capture critical information. To address these limitations, information fusion methods have been proposed, broadly including multi-scale information aggregation [14, 15], difference feature and image feature information aggregation [19], and a combination of both [16]. Unfortunately, these methods tend to introduce noise into the change information. To mitigate this problem, researchers have used methods such as LSTM and RNN to eliminate noise and improve the extraction of global semantic information [20, 22, 26]. Bai [1] and Yang [31] have proposed a change detection method using irregular RNN twin networks that combine Siamese network and RNN. However, these methods suffer from complex structures, high training difficulty, and a high risk of overfitting due to the long-term dependencies of RNN and the large number of LSTM parameters.

In this paper we mitigate the above problem by compensating fine-grained features through CDL.

3 APPROACH

3.1 Overall Framework

In this section we explain our proposed method which addresses the problem of loss of fine-grained information during network downsampling and the difficulty of removing irrelevant change noise in the change detection problem. The framwork of our methods is shown in Fig. 2.

Denote the pair of bitemproal images as $\{X^{t1}, X^{t2}\} \in \mathbb{R}^{C\times H\times W}$, which indicates the past-image and the post-image respectively. And the labelled data is denoted by $Y \in \{0, 1\}$. We define change detection as the process of determining the distance between two distributions, which is considered to have changed when the distance exceeds a certain threshold, i.e. the information represented by the feature has changed. The process can be defined as follows:

$$p = P(\mathcal{D}(\mathcal{F}(X^{t1}), \mathcal{F}(X^{t2}))|X^{t1}, X^{t2}) \tag{1}$$

$$y = \begin{cases} 0, & p \le T \\ 1, & else \end{cases} \tag{2}$$

where P denotes the probabilistic classifier, p denotes the derived probability, $\mathcal{D}$ denotes the distance of the distribution, $\mathcal{F}$ denotes the feature mapping function and T denotes the threshold. Our approach focuses on the noise decoupling of $\mathcal{D}$ process and the features adequate learning among $\mathcal{F}$ process.

Our proposed architecture consists of one backbone for feature extraction, three main components for fine-grained information compensation and noise decoupling, one segmentation head as classifier. Firstly the feature extraction backbone is based on weight-shared Siamese ResNet [10]. Secondly our core method of Fine-grained Information compensation and Noise decOupling (FINO) consists of context-dependent learning (CDL) module, brightness-aware and shape-aware learning (BSA) and regularization gate (REGA) structure. The segmentation head consists of a 1×1 convolution and three 3×3 convolutions in series.

At the beginning, a Siamese network is adopted to map the input images X^{t1}, X^{t2} to low resolution feature tensor. We use features from four stages within the backbone network as representations for downstream change detection. Specifically, we utilize features at resolutions of 1/4, 1/8, 1/16, and 1/32 of the original image resolution. Let $\{X_i^{t1}, X_i^{t2}\}$ denotes the feature maps extracted by the ith stage of the ResNet [10] backbone and C_i is the feature difference of X_i^{t1} and X_i^{t2}.

3.2 Context-Dependent Learning

We claim that the importance of local descriptors within regions far outweighs global descriptors in change detection. Global descriptors mainly manifest as color and brightness changes caused by factors like seasonality and lighting, which can be learned using lightweight global networks. On the other hand, local region features are crucial for decoupling noise between paired images. Based on this understanding, we propose a region-wise attention mechanism that leverages rich semantic features from higher-level network layers to guide lower-level networks in compensating for fine-grained features. In order to obtain the dependencies between image regions, we propose CDL to compensate the fine-grained information which lost during the extraction of high-level semantic

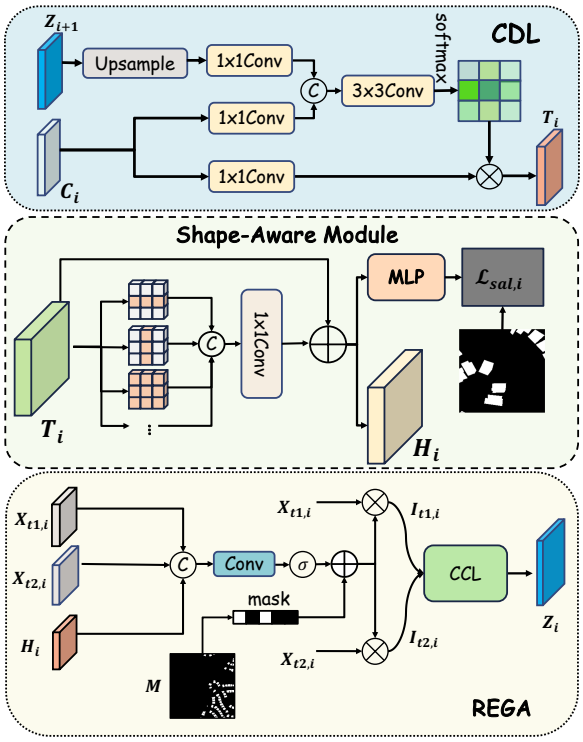

Figure 3: FINO Detail Structure.

features. The CDL consists of multiple region attention modules that are connected in a cascading manner. The architecture is shown in the Fig. 3. To guide the low-level features and fuse the high-level low-noise with the low-level high-noise features, C_i and Z_{i+1} are input to CDL and output M_i, where M_i denotes the changed feature of ith layer and contains both semantic features and fine-grained features.

As the higher level features have stronger semantic information with lower resolution, whereas lower-level network features possess detailed information but at a higher resolution. To apply attention mechanisms efficiently on high-resolution features at minimal cost, we use a localized region attention mechanism instead of the traditional matrix multiplication-based attention mechanism works [2, 5, 14, 23, 25]. This localized attention mechanism focuses on smaller regions within the feature maps, allowing for effective utilization of detailed information without excessive computational overhead. We correct the change features of the current layer by an attention of A_i as follows:

$$A_i(C_i, Z_i) = Softmax(\frac{flatten(\varphi_1(C_i, Z_i))}{\sqrt{dk}}) \tag{3}$$

$$T_i = A_i \varphi_2(C_i) \tag{4}$$

where C_i denotes the difference feature of the Stage-$i-th$ layer in Fig. 3, T_i denotes the region-aware contextual features learned by CDL and φ_1, φ_2 denotes the convolution function.

Since change is often of a localised nature, excessive focus on global attention is not effective. SAM is a local attention that differs from the normal attention. Self-attention uses a attention vector multiplication, while we use element multiplication to compute the attention map.

3.3 Brightness-Aware and Shape-Aware Learning

Typically, due to the large time span of data acquisition for change detection image pairs, objects between image pairs tend to have task agnostic noise. These differences are unpredictable for the change detection task itself and represent global image characteristics. We aim to decouple these noises using brightness-aware and shape-aware modules. And brightness-aware module is only applied at the last layer of the backbone.

Brightness-aware learning. Firstly, the brightness-aware module consists of two contrastive learning components: one for global brightness feature contrastive learning and another for region-level semantic feature alignment. We aim for the backbone network's features to possess robustness against brightness interference in their global features. Therefore, we use learnable global pooling to obtain global features for the paired images and employ contrastive learning to guide the backbone network's learning. The global contrast learning loss is calculated as follows:

$$\mathcal{L}_{gcl} = 1 - \langle \boldsymbol{GAP}(X_{t1}), \boldsymbol{GAP}(X_{t2}) \rangle \tag{5}$$

where $\langle \_, \_ \rangle$ denotes the cosine similarity, $\boldsymbol{GAP}$ indicates the AdaptiveAvgPooling.

Additionally, we aim to minimize the differences between objects in image pairs caused by factors like lighting through feature extraction by the backbone network. Therefore, the loss computation for region-level semantic feature alignment is defined as follows:

$$\mathcal{L}_{rcl} = -\frac{1}{N}\sum_{k}^{N} y_k \log s_k + (1 - y_k) \log(1 - s_k) \tag{6}$$

$$s_k = \langle X_{t1,4,k}, X_{t2,4,k} \rangle \tag{7}$$

where $X_{t1,4,k}, X_{t2,4,k}$ indicates the region feature of forth stage, k is the pixel index in space $N = H \times W$, $\langle \_, \_ \rangle$ denotes the cosine similarity, y_k is the label.

Shape-aware learning. To refine boundary and expand the receptive field, we use a shape-aware module that consists of mutiple asymmetric convolution for anti-aliasing of the edge, shown as in Fig. 3. In our approach, we use seven convolution kernels, including $1 \times 1, 1 \times 3, 3 \times 1, 3 \times 3, 1 \times 5, 5 \times 1, 5 \times 5$. We use Eq to represent the learning of the shape-aware module:

$$H_i = Asym(T_i) \tag{8}$$

$$M_i = MLP(H_i) \tag{9}$$

where $Asym$ indicates the asymmetric convolutions and MLP represents a multilayer perceptron mapping.

Besides, we inject additional supervised objectives in the shape-aware module to ensure the accuracy of estimating shapes. We apply the binary cross-entropy as loss:

$$\mathcal{L}_{sal} = -\sum_{i=1}^{4} \frac{1}{N} \sum_{k}^{N} \bar{y_k} \log M_{i,k} + (1 - y_k) \log(1 - M_{i,k}) \tag{10}$$

where k is the pixel index, $M_{i,k}$ is the i-th stage layer shape mask, y_k is the label.

Additionally, M_i is fed into the regularization gate structure to guide decouple task-specific noise.

3.4 Regularization Gate Structure

We assume that the shape-aware module has obtained highly confident shape features. We use a gating mechanism to selectively enhance foreground features based on this. The specific structure is illustrated in Fig. 3.

Firstly, REGA uses a 1×1 convolution to aggregate features $\{X_{t1,i}, X_{t2,i}\}$ from paired images and shape features H_i learned by asymmetric convolution. Then it passes through the gating mechanism to obtain gating weights. We use sigmoid as the gate. Furthermore, we use the shape mask M_i predicted by the shape-aware module to regularize the gating weights, deactivating neurons in regions affected by pseudo-change noise. It enhances the model robustness to task-specific noise. Finally, a channel attention structure is used for change feature prediction.

The gate function can be defined as Equation 11.

$$G_i = \sigma(\varphi(CAT(X_{t1,i}, X_{t2,i}, H_i))) + M_i I_{t,i} \quad = X_{t,i} \otimes G_i \tag{11}$$

Where $X_{t,i} \in \{X_{t1,i}, X_{t2,i}\}$, $I_{t,i} \in \{I_{t1,i}, I_{t2,i}\}$, i denotes the layer index and $i \in \{1, 2, 3, 4\}$, σ represents the activation function sigmoid, CAT indicates concatenation among features along channel dimension, $\otimes$ indicates spatial-wise multiplication.

Change Characteristics Learning. Existing methods use the distance between two features as the basis for change detection, which lacks interpretability and may miss certain factual change information. As described earlier, we define the procedure of change detection as learning the distance between two distributions. Therefore, we adopt a change characteristics learning module (CCL) based on channel attention to obtain the final change features.

We use the equation 12 and 13 to refer the above operations.

$$d_i = |w \cdot I_i^{t1} - w \cdot I_i^{t2}| \tag{12}$$

$$Z_i = d_i \cdot MLP(MaxPool(d_i) + AvgPool(d_i)) \tag{13}$$

where w denotes the weight that maps the features of bitemproal images to the same distribution space. We then obtain a pixel-by-pixel distribution of distances and input them to the segmentation head to get the prediction.

3.5 Loss

The overall loss function $\mathcal{L}$ consists of change detection loss and supervisory loss:

$$\mathcal{L} = \mathcal{L}_{cd} + \mathcal{L}_{sal} + \lambda \mathcal{L}_{gcl} + \lambda \mathcal{L}_{rcl} \tag{14}$$

where λ indicates a loss weight hyperparameter. We adopt binary cross-entropy for $\mathcal{L}_{cd}$ and $\mathcal{L}_{sal}$.

4 EXPERIMENT

4.1 Settings

Datasets: In this paper we use four CD datasets, named LEVIR-CD [6], WHU [11], DSIFN-CD [32], CDD [12]. We obtain 445/64/128 pairs of patches with size 1024^2 for training/validation/test in LEVIR-CD datasets. We crop the one pair of WHU-CD images to non-overlaped patches with size of 512^2. We obtain 1344/192/384 pairs of patches for training/validation/test. DSIFN-CD is cropped to 256^2 and there are 14400/1360/192 pairs in training/validation/test datasets. The CDD dataset obtain 10000/3000/3000 pairs with size of 256^2 for training/validation/test datasets.

Implementation Details: Our approach is implemented by Pytorch and trained on two GeForce RTX 1080Ti GPUs with 11GB VRAM. The model is trained for 500 epoches with two GPUs, where there are two pairs of images on each GPU. We use ResNet18 as the backbone network in the following experiments. We employ data augmentation techniques, including random rotation, random cropping, brightness variation, and random flipping. The AdamW is used as the optimizer with a *poly* learning rate policy, where the learning rate is set to 0.001 with $\gamma = 0.9$, $weight\_decay = 0.0001$ and policy was step.

Metrics: In change detection, precision reflects resistance to pseudo-change (e.g. color changes of buildings) noise, recall measures sensitivity to actual changes, and F1-score offers a comprehensive evaluation. IoU assesses shape analysis and change object localization. In this context, the primary focus centers on comparing F1-score and IoU.

4.2 Comparison With State-of-the-Art Methods

Comparison of Metrics: In order to demonstrate the effectiveness of our method and compare it with existing excellent CD methods, we conducted comparative experiments with several methods on different datasets. The specific results are shown in Table 1, it can be seen that FINO outperforms other methods and has a significant margin on those datasets. Our method far outperforms the other methods of F1 and IoU on all the four datasets.

The specific results are shown in the Table 3. The Tabel 3 visualizes the considerable improvement of our method over methods such as feature fusion. On both F1 and IoU performing better than DARNet [16] which is the the SOTA, where F1 is 0.43 higher than DARNet and IoU is 0.83 higher than DARNet.

Comparison of Detection Results: In order to demonstrate the superiority of the proposed method in detection results, we conducted a visual analysis on LEVIR-CD and DSIFN-CD. The visual analysis of experimental results are shown in Fig. 4, where the detection results on LEVIR-CD are exhibited in the first row of Fig. 4 (c)-(j) and the detection results on DSIFN-CD are exhibited in the second row of Fig. 4 (c)-(j). In general, FINO performs better than other methods, with smoother and more regular detected edges. For example, as shown in the blue box in Fig. 4(c)-(j), the predictions of FINO are closer to the GT and has smoother and more regular edges, which is more in line with the shape of the building. And FINO has the same sensitivity to large target changes as to small target changes, shown in the red and yellow boxes in Fig. 4(c)-(j). For example, as can be seen from the yellow box, other methods cannot detect this part of the changed object, while FINO is able to not only detect this part of the change, but its predicted label is very close to the GT and its prediction results are more complete and accurate. On the other hand, our method FINO has few errors, as can be observed in the red box.

Comparison with other Feature Fusion and Enhancement methods To compare the superiority of FINO for feature learning at different scales, we compare it with several methods of recent two years that perform well on the LEVIR-CD dataset, which apply

Methods	Backbone	LEVIR-CD [6]				WHU [11]				DSIFN-CD [32]				CDD [12]			
		Pre.(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)	Pre(%)	Rec.(%)	F1(%)	IoU(%)	Pre(%)	Rec(%)	F1(%)	IoU(%)
FC-Siam-diff [17]	U-Net	89.53	83.31	86.31	75.92	47.33	77.66	58.81	41.66	59.67	65.71	62.54	45.50	93.65	54.32	68.76	47.74
SNUNet [9]	U-Net++	89.18	87.17	88.16	78.83	85.60	81.49	83.50	71.67	60.60	72.89	66.18	49.45	95.19	92.51	93.83	88.38
USSFC-Net [13]	U-Net	89.70	92.42	91.04	83.55	89.96	94.56	92.20	85.54	63.73	76.32	69.47	53.21	93.35	96.08	94.74	90.02
IFNet [33]	VGG16	94.02	82.93	88.13	78.77	96.91	73.19	83.40	71.52	67.86	53.94	60.10	42.96	97.64	96.35	96.99	88.94
DASNet [7]	VGG16	90.60	91.38	90.99	83.47	88.23	84.62	86.39	76.04	60.10	56.53	58.26	41.10	92.50	91.40	91.90	-
DTCDSCN [18]	ResNet34	88.53	86.83	87.67	78.05	63.92	82.30	71.95	56.19	53.87	77.99	63.72	46.76	-	-	-	-
TinyCD [8]	EfficientNet	82.68	89.47	91.05	83.57	91.72	91.76	91.74	84.74	51.48	77.66	61.92	44.84	94.78	93.24	94.00	88.68
ScratchFormer [23]	ResNet50	92.70	89.06	90.84	83.22	88.76	84.13	86.39	76.03	88.76	84.14	84.65	73.40	96.29	96.00	96.14	92.57
SARAS-Net [4]	ResNet50	91.97	91.85	91.91	84.95	88.41	85.81	87.09	77.14	67.65	67.51	67.58	51.04	97.76	97.23	97.49	95.11
STANet [21]	ResNet18	83.81	91.00	87.26	70.40	79.37	85.50	82.32	69.95	67.71	61.68	64.56	47.66	95.17	92.88	94.01	87.98
BiT [5]	ResNet18	89.24	89.37	89.31	80.68	86.64	81.48	83.98	72.39	68.36	70.18	69.26	52.97	94.57	88.18	91.26	83.93
Changer [27]	ResNet18	92.86	90.78	91.81	84.86	95.29	89.90	92.49	85.99	88.67	82.20	85.31	74.39	88.67	82.2	85.31	74.39
STNet [21]	ResNet18	92.06	89.03	90.52	82.09	87.84	87.08	87.46	77.72	-	-	-	-	-	-	-	-
APD [30]	ResNet18	92.81	90.64	91.71	84.69	95.74	86.94	91.13	83.70	89.39	86.40	87.87	78.36	96.36	96.01	96.20	92.67
CANet [34]	ResNet18	92.52	90.06	91.27	83.95	-	-	-	-	-	-	-	-	96.82	97.29	97.05	94.27
FINO(Ours)	ResNet18	93.70	91.15	92.41	85.89	95.27	97.26	93.35	90.96	91.34	96.01	89.35	80.75	98.76	98.57	98.66	97.36

Table 1: Quantitative results on LEVIR-CD dataset, WHU-CD dataset DSIFN-CD dataset and CDD dataset. The best values are in red and the second are in blue. We add references of these models in the supplementary material.

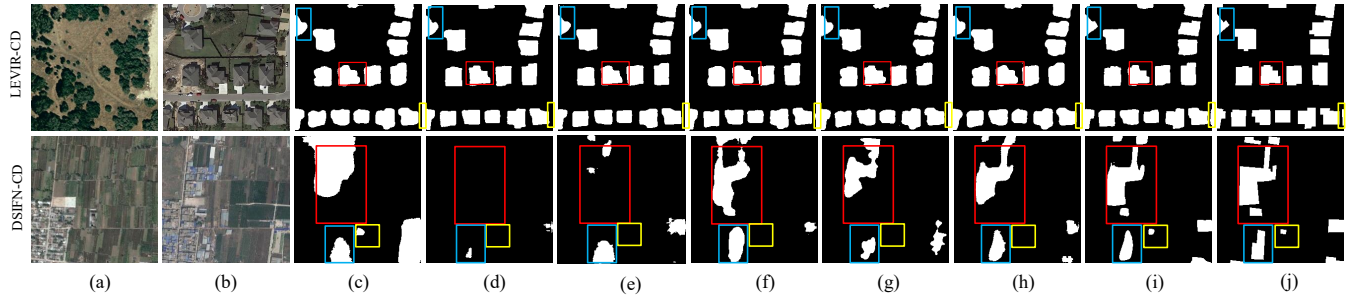

Figure 4: Qualitative results of different CD methods on LEVIR-CD. (a) T_1 image. (b) T_2 image. (c) FC-EF. (d) FC-Siam-Diff. (e) FC-Siam-Conc. (f) DTCDSCN. (g) BIT. (h) ChangeFormer. (i) FINO(ours). (j) Ground truth. where the first row shows the comparison of detection results for dense objects and the second row shows the comparison of detection results for multi-scale objects. The yellow box is the detection effect of small objects, the red box is the detection effect of large objects, and the blue box is the detection effect of object edges.

Methods	Pre.(%)	Rec.(%)	F1(%)	IoU(%)
RSCD [15]	92.15	89.73	90.93	83.36
D-TNet18 [29]	89.13	88.58	88.85	79.94
DARNet [16]	92.67	91.31	91.98	85.16
ChangeFormer [2]	92.05	88.80	90.40	82.48
DESSN [14]	90.99	91.73	91.36	-
FINO(Ours)	93.70	91.15	92.41	85.89

Table 2: Comparison experiments of our method with other direct feature enhancement or feature fusion methods on LEVIR-CD dataset.

Methods	Pre.(%)	Rec.(%)	F1(%)	IoU(%)
RSCD	92.15	89.73	90.93	83.36
D-TNet18	89.13	88.58	88.85	79.94
DARNet	92.67	91.31	91.98	85.16
ChangeFormer	92.05	88.80	90.40	82.48
DESSN	90.99	91.73	91.36	-
FINO(Ours)	93.70	91.15	92.41	85.89

Table 3: Comparison experiments of our method with other direct feature enhancement or feature fusion methods on LEVIR-CD dataset.

feature concatenation [15], feature interaction [14, 16], skip connection [29], and attention mechanisms [2] to achieve feature fusion of different scales.

4.3 Comparison with other Feature Fusion and Enhancement methods

To compare the superiority of FINO for feature learning at different scales, we compare it with several methods of recent two years that perform well on the LEVIR-CD dataset, which apply feature concatenation [15], feature interaction [14, 16], skip connection [29], and attention mechanisms [2] to achieve feature fusion of different scales.

The specific results are shown in the Table 3. The Tabel 3 visualizes the considerable improvement of our method over methods such as feature fusion. On both F1 and IoU performing better than DARNet [16] which is the the SOTA, where F1 is 0.43 higher than DARNet and IoU is 0.83 higher than DARNet.

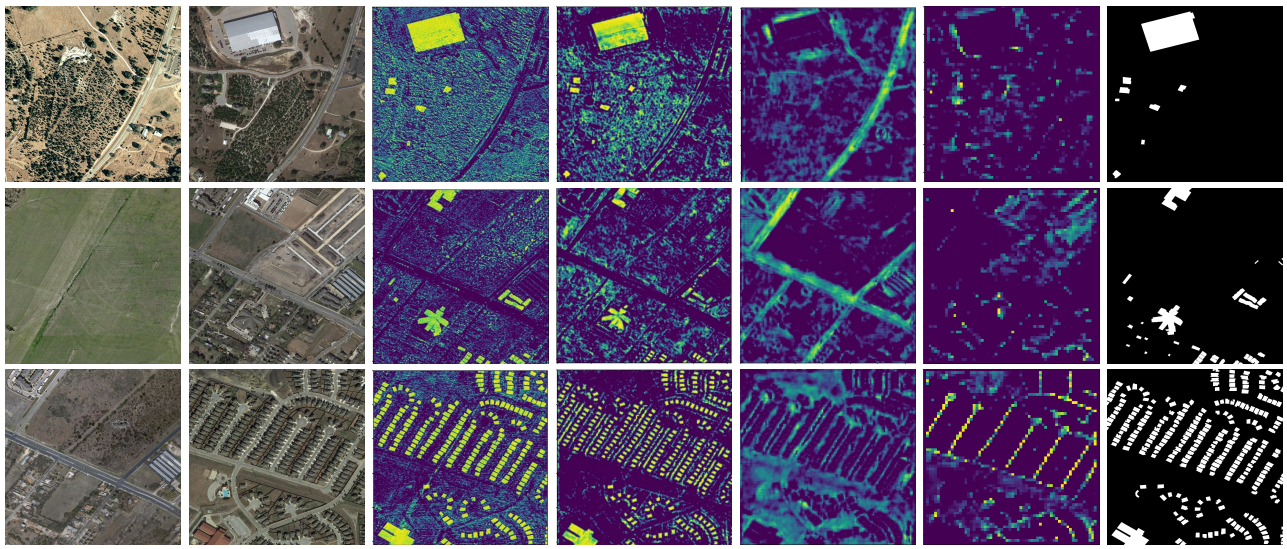

Figure 5: Feature visualisation on LEVIR-CD. (a) T_1 image. (b) T_2 image. (c) change features C_1 of Stage-1. (d) change features C_2 of Stage-2. (e) change features C_3 of Stage-3. (f) change features C_4 of Stage-4. (g) Ground Truth. Yellow pixels represent higher probability of change and blue pixels represent lower probability of unchanged.

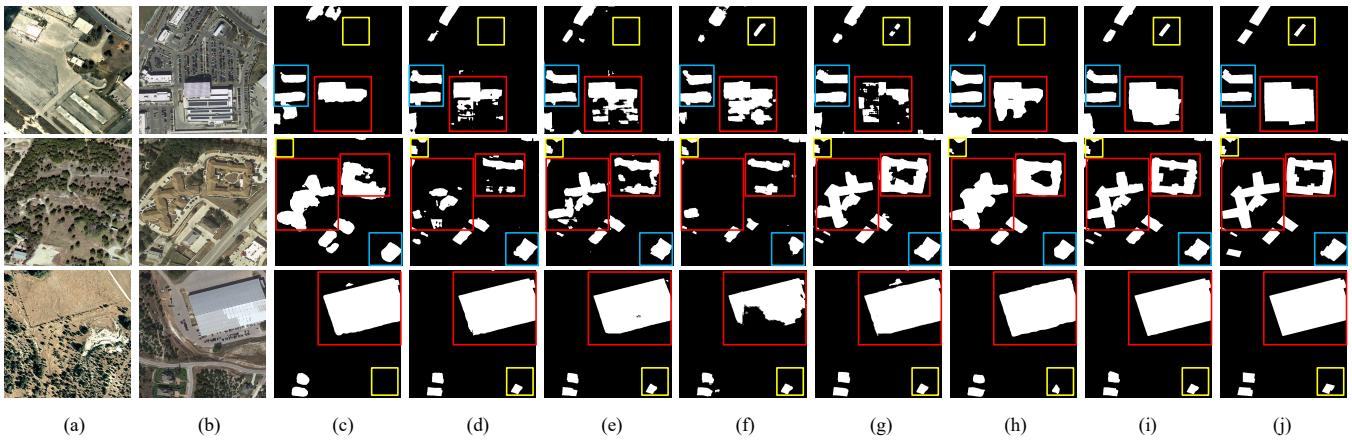

Figure 6: Qualitative results of different CD methods of targets at different scales. (a) T_1 image. (b) T_2 image. (c) FC-Siam-diff. (d) BiT. (e) ChangeFormer (f) USSFC-Net. (g) SARAS-Net. (h) Changer. (i) FINO(ours). (j) Ground truth.

4.4 Detection Result

To strongly support our methods, we compare the detection effects in different contexts on LEVIR-CD. Overall from the whole images, our method outperforms the other methods. Specifically, as shown in Figure 6, our method has excellent detection results for both large targets labelled by red boxes, medium targets labelled by blue boxes and small targets labelled by yellow boxes in the same scene.

As shown in Figure 6, our method also performs well with dense targets. As described in the main text, the context perceptual reconstruction architecture (FINO) architecture is designed to recover the detailed information such as scale and shape. Therefore, our method is more sensitive to the scale and shape of the target. As shown in Figure 6, the boundaries between the dense targets are very clear in our result, unlike other methods that miss some targets or have fuzzy boundaries. Especially in the regions we have marked with red boxes. Due to the high resolution, please zoom in to view the image if necessary, it will be clearer.

As shown in Figure 6, our method also outperforms other methods in complex scenes. Our method is more robust to noise and provides more accurate detection results. The buildings are highly similar to their surroundings, as shown in the first and second rows of the Figure 6. And the buildings are very similar in shape and

Replace	Pre.(%)	Rec.(%)	F1(%)	IoU(%)
Concatenation	92.53	89.79	91.14	83.72
Cross-attention	88.85	81.37	84.94	73.82
No-local attention	90.80	88.64	89.75	81.41
SPP attention	91.97	89.44	90.69	82.96
CDL	93.70	91.15	92.41	85.89

Table 4: Comparison of CD performance on LEVIR-CD dataset by using our proposed CDL and other attention methods.

stages	Pre.(%)	Rec.(%)	F1(%)	IoU(%)
4	87.43	86.84	87.84	78.31
4,3	92.00	89.58	90.77	83.11
4,3,2	93.45	90.47	91.93	85.07
4,3,2,1	93.70	91.15	92.41	85.89

Table 5: A comparative experiment on LEVIR-CD of stages in the cascade structure of our approach, where stages represent the location number of the ResNet network layers in Fig. 2 that used FINO.

colour to the surrounding bush noise, especially as shown in the the red boxes in the second rows and the third rows. Comparing columns (c), (h), and (i), it is clear that our method performs better. Comparing columns (d)-(g) and (i), there are few targets missed. However they detect some noise as a change (e.g. the blue box) because they are oversensitive about the change. Since feature elimination and aggregation module (FEAM) is designed to eliminate noise and robustly enhance the foreground, our method detects all the changed targets without being disturbed by these noises. This demonstrates the high robustness of our approach.

4.5 Ablation Study and Further Analysis

In this section, we provide ablation studies and an analysis of our model. Particularly, we explore the contributions of main components.

4.5.1 Importance of the Main Components. To verify the usefulness of our FINO for information recovery, we conducted ablation experiments for the positions of the cascade structure, where the network layers are numbered 1, 2, 3 and 4 from the lower to the higher levels in Fig. 2. The specific results are shown in Table 5. As can be seen in Table 5, the accuracy of change detection are greatly improved as the more layers employed FINO. In particular, during the process from Stage-4 to Stage-2, the improvement of FINO is significant.

To verify the effectiveness of each component, ablation experiments are carried out on CDL, BSA and REGA, as shown in the Table 6. We use Siam-ResNet as a baseline and verify the performance improvements of the above components on the method in turn. As can be observed from the F1 and IoU in the second row of the Table 6, the CDL improves the performance of the method

Row	CDL	BSA	REGA	Pre.(%)	Rec.(%)	F1(%)	IoU(%)
1				87.26	86.70	87.02	77.03
2	√			90.14	89.73	89.93	81.71
3		√		92.22	90.56	91.38	84.13
4	√	√		93.14	90.27	91.68	84.64
5	√	√	√	93.70	91.15	92.41	85.89

Table 6: Object change detection experimental results on LEVIR-CD for validation of the effectiveness of each component, where Sup. denotes the indirect intermediate supervision of learnable mask.

enormously that further increase F1 by 2.91% and IoU by 4.68%. It can be noticed from the first and third row of the table that the task-specific nois decouple module REGA is able to improve the model performance greatly, making the F1 increase by 4.36% and IoU by 7.1%. As shown in the first and forth row of Table 6, F1 increases 4.66% and IoU increases 7.61%. The BSA further enhances our capability to detect changes in our approach, shown in row 5 of Table 6.

4.5.2 Importance of Fine-Grained Information. To verify the effectiveness of CDL and proof of significance of fine-grained information, we replace CDL with several other types of attention, including feature concatenation, cross-attention, no-local attention and spatial pyramid pooling (SPP) attention. And due to space constraints, specific structural details are provided in the supplementary material.

The results on LEVIR-CD are shown in Table 4. Since other attention mechanisms have to rely on matrix multiplication, which has high computational complexity at high resolution, we have to downsample. We downsample the features to a size of 64×64. As shown in Table 4, our CDL has a significant superiority over other attention mechanisms, specifically, CDL improves the F1 metric by 1.27% and IoU by 2.17% over the Concatenation, which is the best among other attention mechanisms. This further validates the importance of the importance of fine-grained information in change detection.

Visualization of FINO In order to objectively show whether the fine-grained information is adequately compensated by FINO, we visualize the difference features of the four stages in FINO. It is shown in Fig. 5. In the feature maps of column (c)-(d) in Fig. 5, the pixels around the yellow areas appear darker and have a clear contrast with the surrounding pixels, with details of objects clearly. It proves that FINO is effective in noise elimination in non-changing regions and can effectively recover detailed texture information of images. In the feature maps column (e)-(f) of Fig. 5, no details such as the shape of objects can be distinguished, but the outline is clear. This proves that as the network layer deepens, showing in column (e)-(f) of Fig. 5, the high-level semantic features of object are very obvious and less noisy, but lost a lot of detail informations. And due to FINO, the low-level change features of Fig. 5 (c) has few noise, including the higher-level semantics of the change features retained, which is used to predict changes. And comparing with

the GT of Fig. 5 (g), all the yellow region in the features of Fig. 5 (c) are the change objects.

5 CONCLUSION

In this article, we propose the fine-grained information compensation and noise decoupling (FINO) architecture to address the issues of information loss in convolutional networks and the interference of pseudo-change noise in change detection. Our designed context-aware learning (CDL) fully compensates for fine-grained information. We introduce a brightness-aware and shape-aware module to mitigate task-agnostic noise, and propose a regularization gate (REGA) structure to decouple task-specific noise. Extensive experiments demonstrate the advantages of our FINO, outperforming current state-of-the-art methods on mutiple change detection datasets.

REFERENCES

[1] Beifang Bai, Wei Fu, Ting Lu, and Shutao Li. 2022. Edge-Guided Recurrent Convolutional Neural Network for Multitemporal Remote Sensing Image Building Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–13. https://doi.org/10.1109/TGRS.2021.3106697

[2] Wele Gedara Chaminda Bandara and Vishal M. Patel. 2022. A Transformer-Based Siamese Network for Change Detection. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. 207–210. https://doi.org/10.1109/IGARSS46834.2022.9883686

[3] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. 2018. Fully Convolutional Siamese Networks for Change Detection. (2018), 4063–4067. https://doi.org/10.1109/ICIP.2018.8451652

[4] Chao-Peng Chen, Jun-Wei Hsieh, Ping-Yang Chen, YI-Kuan Hsieh, and Bor-Shiun Wang. 2023. SARAS-Net: Scale and Relation Aware Siamese Network for Change Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (Jun. 2023), 14187–14195. https://doi.org/10.1609/aaai.v37i12.26660

[5] Hao Chen, Zipeng Qi, and Zhenwei Shi. 2022. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14. https://doi.org/10.1109/TGRS.2021.3095166

[6] Hao Chen and Zhenwei Shi. 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing* 12, 10 (2020). https://doi.org/10.3390/rs12101662

[7] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. 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 (2021), 1194–1206. https://doi.org/10.1109/JSTARS.2020.3037893

[8] Andrea Codegoni, Gabriele Lombardi, and Alessandro Ferrari. 2022. TINYCD: A (Not So) Deep Learning Model For Change Detection. *arXiv preprint arXiv:2207.13159* (2022).

[9] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. 2022. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5. https://doi.org/10.1109/LGRS.2021.3056416

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[11] Shunping Ji, Shiqing Wei, and Meng Lu. 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 (2019), 574–586. https://doi.org/10.1109/TGRS.2018.2858817

[12] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis. 2018. CHANGE DETECTION IN REMOTE SENSING IMAGES USING CONDITIONAL ADVERSARIAL NETWORKS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2 (2018), 565–571. https://doi.org/10.5194/isprs-archives-XLII-2-565-2018

[13] Tao Lei, Xinzhe Geng, Hailong Ning, Zhiyong Lv, Maoguo Gong, Yaochu Jin, and Asoke K. Nandi. 2023. Ultralightweight Spatial–Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14. https://doi.org/10.1109/TGRS.2023.3261273

[14] Tao Lei, Jie Wang, Hailong Ning, Xingwu Wang, Dinghua Xue, Qi Wang, and Asoke K. Nandi. 2022. Difference Enhancement and Spatial–Spectral Nonlocal Network for Change Detection in VHR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–13. https://doi.org/10.1109/TGRS.2021.3134691

[15] Zhenglai Li, Chang Tang, Lizhe Wang, and Albert Y. Zomaya. 2022. Remote Sensing Change Detection via Temporal Feature Interaction and Guided Refinement. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–11. https://doi.org/10.1109/TGRS.2022.3199502

[16] Ziming Li, Chenxi Yan, Ying Sun, and Qinchuan Xin. 2022. A Densely Attentive Refinement Network for Change Detection Based on Very-High-Resolution Bitemporal Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–18. https://doi.org/10.1109/TGRS.2022.3159544

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 936–944. https://doi.org/10.1109/CVPR.2017.106

[18] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. 2021. 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 (2021), 811–815. https://doi.org/10.1109/LGRS.2020.2988032

[19] Zhiyong Lv, Fengjun Wang, Guoqing Cui, Jón Atli Benediktsson, Tao Lei, and Weiwei Sun. 2022. Spatial–Spectral Attention Network Guided With Change Magnitude Image for Land Cover Change Detection Using Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–12. https://doi.org/10.1109/TGRS.2022.3197901

[20] Haobo Lyu and Hui Lu. 2016. Learning a transferable change detection method by Recurrent Neural Network. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. https://doi.org/10.1109/IGARSS.2016.7730344

[21] Xiaowen Ma, Jiawei Yang, Tingfeng Hong, Mengting Ma, Ziyan Zhao, Tian Feng, and Wei Zhang. 2023. STNet: Spatial and Temporal feature fusion network for change detection in remote sensing images. *arXiv preprint arXiv:2304.11422* (2023).

[22] Lichao Mou, Lorenzo Bruzzone, and Xiao Xiang Zhu. 2019. Learning Spectral-Spatial-Temporal Features via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing* 57, 2 (2019), 924–935. https://doi.org/10.1109/TGRS.2018.2863224

[23] Mubashir Noman, Mustansar Fiaz, Hisham Cholakkal, Sanath Narayan, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. 2022. Remote Sensing Change Detection with Transformers Trained from Scratch. *arXiv preprint arXiv:2304.06710*.

[24] Jinlong Peng, Zhengkai Jiang, Yueyang Gu, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, and Weiyao Lin. 2021. Siamrcr: Reciprocal classification and regression for visual object tracking. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[25] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. 2020. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European Conference on Computer Vision (ECCV)*.

[26] Marc Rußwurm and Marco Körner. 2017. Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1496–1504. https://doi.org/10.1109/CVPRW.2017.193

[27] Zhe Li Sheng Fang, Kaiyu Li. 2023. Changer Feature Interaction is What You Need for Change Detection. *Arxiv* https://doi.org/10.48550/arXiv.2209.08290 (2023).

[28] Ashley Varghese, Jayavardhana Gubbi, Akshaya Ramaswamy, and P. Balamuralidhar. 2019. ChangeNet: A Deep Learning Architecture for Visual Change Detection. In *Computer Vision – ECCV 2018 Workshops*, Laura Leal-Taixé and Stefan Roth (Eds.). Springer International Publishing, Cham, 129–145.

[29] Ling Wan, Ye Tian, Wenchao Kang, and Lei Ma. 2022. D-TNet: Category-Awareness Based Difference-Threshold Alternative Learning Network for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–16. https://doi.org/10.1109/TGRS.2022.3213925

[30] Supeng Wang, Yuxi Li, Ming Xie, Mingmin Chi, Yabiao Wang, Chengjie Wang, and Wenbing Zhu. 2023. Align, Perturb and Decouple: Toward Better Leverage of Difference Information for RSI Change Detection. arXiv:2305.18714 [cs.CV]

[31] Bin Yang, Le Qin, Jianqiang Liu, and Xinxin Liu. 2022. IRCNN: An Irregular-Time-Distanced Recurrent Convolutional Neural Network for Change Detection in Satellite Time Series. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5. https://doi.org/10.1109/LGRS.2022.3154894

[32] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. 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 (2020), 183–200. https://doi.org/10.1016/j.isprsjprs.2020.06.003

[33] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. 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 (2020), 183–200.

[34] Feng Zhou, Chao Xu, Renlong Hang, Rui Zhang, and Qingshan Liu. 2023. Mining Joint Intraimage and Interimage Context for Remote Sensing Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–12.