Align, Perturb and Decouple: Toward Better Leverage of Difference Information for RSI Change Detection

Supeng Wang[1], **Yuxi Li**[2], **Ming Xie**[1], **Mingmin Chi**[1,3,4*], **Yabiao Wang**[2*], **Chengjie Wang**[2,5*], **Wenbing Zhu**[6]

[1]School of Computer Science, Shanghai Key Laboratory of Data Science, Fudan University
[2]YouTu Lab, Tencent, Shanghai
[3]Zhengzhou Zhongke Institute of Integrated circuit and System Application, China
[4]Zhongshan PoolNet Technology Ltd, China
[5]Shanghai Jiao Tong University, China
[6]Rongcheer, China
spwang21@m.fudan.edu.cn, {mxie20,mmchi}@fudan.edu.cn,
{yukiyxli,caseywang,jasoncjwang}@tencent.com, louis.zhu@rongcheer.com

Abstract

Change detection is a widely adopted technique in remote sense imagery (RSI) analysis in the discovery of long-term geomorphic evolution. To highlight the areas of semantic changes, previous effort mostly pays attention to learning representative feature descriptors of a single image, while the difference information is either modeled with simple difference operations or implicitly embedded via feature interactions. Nevertheless, such difference modeling can be noisy since it suffers from non-semantic changes and lacks explicit guidance from image content or context. In this paper, we revisit the importance of feature difference for change detection in RSI, and propose a series of operations to fully exploit the difference information: Alignment, Perturbation and Decoupling (APD). Firstly, alignment leverages contextual similarity to compensate for the non-semantic difference in feature space. Next, a difference module trained with semantic-wise perturbation is adopted to learn more generalized change estimators, which reversely bootstraps feature extraction and prediction. Finally, a decoupled dual-decoder structure is designed to predict semantic changes in both content-aware and content-agnostic manners. Extensive experiments are conducted on benchmarks of LEVIR-CD, WHU-CD and DSIFN-CD, demonstrating our proposed operations bring significant improvement and achieve competitive results under similar comparative conditions. Code is available at https://github.com/wangsp1999/CD-Research/tree/main/openAPD

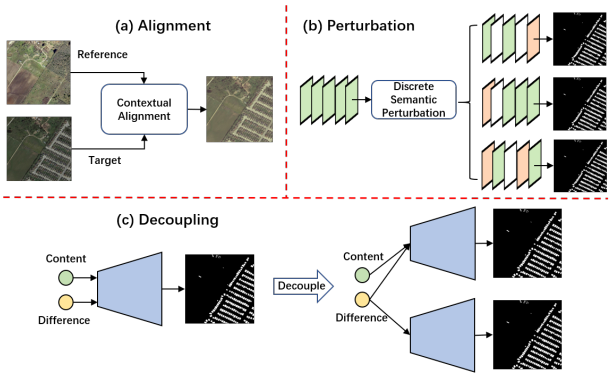

Figure 1: Conceptual illustration of operations proposed in this work to improve utilization of difference information. (a) A context reliant alignment operation to mitigate changes irrelevant to semantic. (b) A difference guidance module trained with discrete semantic perturbation. (c) Traditional single decoder with hybrid input is decoupled into a content-aware decoder and content agnostic decoder.

1 Introduction

Change detection is a vision task aiming at identifying pixel-wise semantic changes between paired bitemporal images, this technique is helpful for analysis of remote sense imagery (RSI) with high resolution, which provides important information about changes in land surface and expansion of settlement during long time of observation [Hussain *et al.*, 2013; Daudt *et al.*, 2018b].

With the development of remote satellite sensoring, there exist more open source databases of remote sense imagery with fine-grained semantic annotations [Chen and Shi, 2020a; Ji *et al.*, 2018; Zhang *et al.*, 2020a], which makes it possible to exploit the data-hungry deep-learning approaches [Krizhevsky *et al.*, 2017; He *et al.*, 2016] to achieve more accurate change detection. Due to the nature of pair-wise input and dense prediction, encoder-decoder ar-

*Corresponding author

chitectures with siamese backbones prevail in recent efforts, where features of single images are extracted separately and the difference information is exploited in a decoder to highlight areas of changing objects [Daudt *et al.*, 2018a; Bandara and Patel, 2022; Chen *et al.*, 2022b; Zheng *et al.*, 2021; Chen *et al.*, 2022a].

In a nutshell, recent approaches usually model the difference information between input pairs in simple and direct manners (e.g. taking difference or concatenation to obtain difference information) [Daudt *et al.*, 2018a; Bandara and Patel, 2022; Zheng *et al.*, 2021] or implicitly embed the difference into feature interaction [Chen *et al.*, 2022b; Chen *et al.*, 2022a], while still leaving some key issues open. **Firstly**, difference information is inherently vulnerable to pseudo-changes (e.g. the seasonal illumination changes during long-term observation), but few of previous works explicitly take such interference into account, hence these methods are not guaranteed to be robust enough to non-semantic changes. **Secondly**, most of previous literature take difference information from features for final decoding, ignoring the fact that difference naturally contains information of changeful objects, which reversely provides spatial guidance to representation learning [Hussain *et al.*, 2013]. **Finally**, the relationship between image content and difference information is seldom discussed in prior works, the image content can be regarded as auxiliary prior information for change decoding, while also introducing some irrelevant change cues distracting prediction.

With the reviews above, we claim that the difference information in current research is still underutilized. Therefore, in this paper, we design series of operations aiming at mitigating the aforementioned issues and fully leveraging feature difference to boots change detection results. Concretely, we equip the hierarchical encoder-decoder network with three operations in sequential order: **A**lignment, **P**erturbation and **D**ecoupling (APD), which is illustrated in Figure 1. **Alignment:** To alleviate noise from pseudo-changes, we first propose a graph-based alignment module, which exploits the contextual similarity between patches to aggregate information from areas of the same semantic as compensation, this results in more precise extraction of semantic difference in following stages. **Perturbation:** Following the alignment operation, we propose a perturbation-aided difference prediction module. Especially, this module is trained with discrete semantic perturbation as feature augmentation, thus can recognize more generalized change patterns as guidance for feature extraction in the following stages. **Decoupling:** We decouple the decoder into an asymmetric dual-stream structure for final prediction, one focuses on integrating bitemporal contents with difference while the other takes pure difference information for decoding, this helps utilize the complementary property between image content and difference information while avoiding irrelevant noise.

We conduct experiments on three challenging benchmarks for RSI change detection: LEVIR-CD [Chen and Shi, 2020a], WHU-CD [Ji *et al.*, 2018] and DSIFN-CD [Zhang *et al.*, 2020a], and demonstrate the superiority of our proposed approach over existing competitive methods. Plenty of ablation studies also verify the effectiveness of our proposed operations. In a nutshell, the contribution of this paper can be summarized as following

- We reconsider the problem of pseudo-changes and propose a graph-based alignment module to explicitly utilize contextual similarity for compensation.
- We propose a hierarchical difference extraction structure to guide the feature extraction process stage-by-stage, which is equipped with a specially designed discrete semantic perturbation scheme for feature augmentation.
- Different from most of prior works, we propose a dual decoder structure to decouple the utilization of image content from pure difference encoders.
- We integrate the proposed operation above and convert traditional encoder-decoder structure into a new change detector APD, which achieves competitive results on mainstream benchmarks.

2 Related Works

2.1 Change Detection with Deep Learning

We roughly divide deep learning-based change detection methods into two types [Zhang and Shi, 2020]: two-stage and one-stage methods. In general, the two-stage method trains a CNN/FCN to classify the bitemporal images respectively, and compares their classification results to obtain change areas. To achieve this goal, both the bitemporal semantic labels and the change label should be provided [Nemoto *et al.*, 2017; Liu *et al.*, 2019a; Ji *et al.*, 2019].

The one-stage method is a more prevailed framework in recent research, which takes change detection as a dense classification task and directly produces the change result from the bitemporal images. The FC series [Daudt *et al.*, 2018a] was one of the earliest methods to adopt convolution neural networks for change detection, where three architectures were proposed: FC-EF, FC-Siam-Conc, and FC-Siam-Diff. FC-EF adopted the early fusion strategy, while FC-Siam-Conc and FC-Siam-Diff adopted the medium fusion strategy with different difference policies. Besides, DTCDSCN [Liu *et al.*, 2019b] takes inspiration from the two-stage methods, and converts change detection to multi-task pipeline with semantic map as auxiliary supervision. ChangeSTAR [Zheng *et al.*, 2021] further relaxes the paired requirement by taking unpaired images as input.

2.2 Attention Mechanism in Change Detection

Although deep neural network achieves significant improvement over handcrafted features, traditional network structures can not explicitly capture the contextual reliance within RSI, thus can not accurately detect changes in object-level. Therefore, there appears works resorting to attention mechanism [Jaderberg *et al.*, 2015; Hu *et al.*, 2017; Vaswani *et al.*, 2017] to help gather more contextual information. Zhang et al.[Zhang *et al.*, 2020b] proposed a deeply supervised change detection network (IFN) , which applies channel attention and spatial attention for feature enhancement.

Besides the contextual modeling, recent works also make attempt to integrate the attention mechanism into bitemporal interaction as an implicit replacement of feature differ-

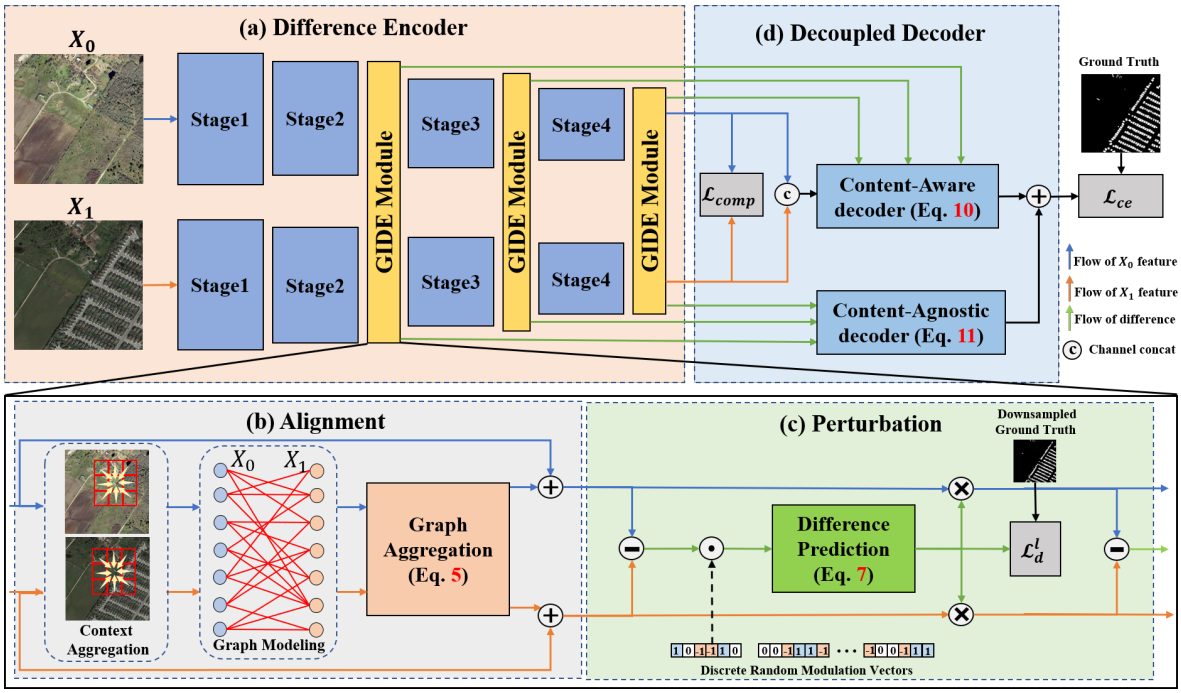

Figure 2: The overall framework of our proposed APD change detector. (a) The part of the encoder equipped with our proposed GIDE Module. (b) Detailed illustration of graph-based alignment within GIDE. (c) Illustration of the perturbation guided difference prediction in GIDE. (d) The proposed asymmtetrical duel decoder part for final prediction.

ence operation. Both BIT [Chen *et al.*, 2022b] and TransUNetCD [Li *et al.*, 2022] proposed a bitemporal image transformer by mixing CNN and transformer, where the cross image interaction is conducted in latent encoded space. ChangeFormer [Bandara and Patel, 2022] further developed a siamese variant of SegFormer [Xie *et al.*, 2021] for feature extraction.

2.3 Feature Augmentation

Feature augmentation aims at perturbing data in the encoded feature space to help the neural network learn invariant representation. Different from data augmentation with low-level transformation, feature-level augmentation is regarded as increasing the semantic variance to ensure the generalization ability of models. ISDA [Wang *et al.*, 2019] perturbed the deep features in terms of their statistical characteristic, thus improving the accuracy of recognition. The work of SFA [Li *et al.*, 2021] utilized a similar adaptive noise for domain generalization. Further, Manifold Mixup [Verma *et al.*, 2019] extended the widely-used mixup trick into feature space to smooth the semantic boundaries. MGD [Yang *et al.*, 2022] introduced random mask for knowledge distillation among deep features to obtain better results of student models.

In change detection, the difference information is usually derived from the discrepancy between two feature maps, which should be invariant to the difference order and channel-wise masking. Therefore, we take inspiration from feature augmentation methods and design a specialized discrete semantic perturbation mechanism to help difference prediction module recognize more general patterns.

3 Approach

Figure 2 depicts the pipeline of our APD change detector. Similar to previous work, our method is built based on an encoder-decoder architecture. We denote the bitemporal image pair as $\mathcal{X}_0$ and $\mathcal{X}_1 \in \mathbb{R}^{3\times H\times W}$ with change label map $\mathcal{Y} \in \{0,1\}^{H\times W}$, then feed images to a classic hierarchical deep neural network to extract feature representations of multi-level $\{\mathcal{F}_0^l, \mathcal{F}_1^l\}_l$, where $l \in [2, N]$ represents the stage index (at most N stages). However, to fully leverage the difference information, we propose the concept of "Align First and Then Difference", and insert a **G**raph **I**nteractive and **D**ifference **E**nhancement module (GIDE) between stages in the encoder part, which transforms $\mathcal{F}_i^l$ into more change-sensitive manifold $\widetilde{\mathcal{F}}_i^l$

$$\begin{aligned} \widetilde{\mathcal{F}}_0^l, \widetilde{\mathcal{F}}_1^l, \mathcal{O}^l &= \textbf{GIDE}(\mathcal{F}_0^l, \mathcal{F}_1^l) = P\left(A\left(\mathcal{F}_0^l, \mathcal{F}_1^l\right)\right) \\ \mathcal{F}_0^{l+1}, \mathcal{F}_1^{l+1} &= \textbf{Enc}_l\left(\widetilde{\mathcal{F}}_0^l, \widetilde{\mathcal{F}}_1^l\right), \end{aligned} \tag{1}$$

where $\textbf{Enc}_l(\cdot,\cdot)$ denotes the l-th stage of the encoder. Specifically, GIDE is composed of an alignment module $A(\cdot,\cdot)$, which aggregates pixel-level semantic information, and a perturbation-aided difference module $P(\cdot,\cdot)$ which augments feature difference and highlights local areas. Additional to the alignment and perturbation, $P(\cdot,\cdot)$ also generates feature difference $\mathcal{O}^l$ as pure difference information for the following decoding. In the part of the decoder, we decouple the classical single decoder into a complementary structure of content-aware decoder D_{aw} and content-agnostic decoder D_{ag} to excavate the content and difference information respectively for final prediction.

3.1 Context-aware Alignment

The alignment operation aims at alleviating the disturbance of pseudo-changes and amplifying the area of semantic change, thus we resort to an undirected bipartite graph as an aggregation structure to align features (as shown in Figure 2 (b)).

Formally, suppose both $\mathcal{F}_0^l, \mathcal{F}_1^l$ are of resolution $H_l \times W_l$, then we can take each pixel in feature map as a graph node and build an adjacent matrix $\mathcal{G}$, which is a bipartite graph, i.e. an edge can only link pixels from different feature maps, thus the adjacent matrix can be represented as four-dimensional, $\mathcal{G} \in \{0,1\}^{H_l \times W_l \times H_l \times W_l}$, and the node feature is the feature of the corresponding pixel. Considering the feature of a single pixel can be sensitive to non-semantic changes, while its contextual information can be more robust since it contains more structural information, therefore it is more suitable to encode the edge between nodes with the reference of contextual similarity. To do this, we first obtain the context feature $\mathcal{F}_{0,c}^l, \mathcal{F}_{1,c}^l$ via a non-parameterized dilated convolution of dilate factor d and kernel size k, for $m = 0, 1$, we have

$$\mathcal{F}_{m,c}^l[u,v] = \sum_{\substack{(i,j)\in[-k,k]^2 \\ (i,j)\neq(0,0)}} \mathcal{F}_m^l[u+id, v+jd], \tag{2}$$

where $[u, v]$ indicates obtaining the pixel-wise feature at location (u, v). To build graph $\mathcal{G}$, for each pixel (u, v) on $\mathcal{F}_{0,c}^l$, we compute top n nearest coordinates on $\mathcal{F}_{1,c}^l$

$$\mathcal{H}_{u,v} = \mathbf{TopK}_{i\in[0,H_l],j\in[0,W_l]}\left(\left\|\mathcal{F}_{0,c}^l[u,v] - \mathcal{F}_{1,c}^l[i,j]\right\|_2\right), \tag{3}$$

with the grouped coordinate set of nearest neighbor $\mathcal{H}_{u,v}$, the high-dimensional adjacent matrix can be expressed as

$$\mathcal{G}[u,v,i,j] = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{H}_{u,v} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

With the relation graph $\mathcal{G}$ based on context, we can take the naive graph convolution to aggregate information from similar node pairs to both feature maps as follows

$$\begin{aligned} \mathcal{F}_0^{l\prime}, \mathcal{F}_1^{l\prime} &= \mathbf{GCN}(\mathcal{G}, \mathcal{F}_0^l, \mathcal{F}_1^l) \\ \hat{\mathcal{F}}_0^l = \mathcal{F}_0^l + \mathcal{F}_0^{l\prime} &\quad \hat{\mathcal{F}}_1^l = \mathcal{F}_1^l + \mathcal{F}_1^{l\prime}, \end{aligned} \tag{5}$$

where $\hat{\mathcal{F}}_0^l, \hat{\mathcal{F}}_1^l$ are aligned features after aggregation of graph convolution and will be utilized in following difference prediction module.

3.2 Perturbation-aided Difference Prediction

After the feature alignment, we make attempts to leverage the aligned feature to produce coarse spatial guidance, thus helping feature extraction in the following stage. To this end, we introduce a feature augmentation mechanism to train a coarse difference prediction module (as described in Figure 2 (c)).

We start with a simple difference operation, however, we explicitly modulate the channel of feature difference with a random vector $\mathbf{v}$ filled with discrete values

$$\begin{aligned} \hat{\mathcal{O}}^l &= (\hat{\mathcal{F}}_0^l - \hat{\mathcal{F}}_1^l) \odot \mathbf{v} \\ \textbf{s.t.} \quad \|\mathbf{v}\|_1 &= (1-\tau)C^l \quad \mathbf{v} \in \{+1, -1, 0\}^{C^l}, \end{aligned} \tag{6}$$

where C^l is the channel dimension of $\hat{\mathcal{F}}_0^l, \hat{\mathcal{F}}_1^l$, and $\odot$ represents channel-wise modulation, i.e. the i-th element of vector $\mathbf{v}$ is broadcasted and multiplied to all pixels on the i-th channel. The value of vector $\mathbf{v}$ is constrained within 0 and ± 1, thus each channel of feature difference is either reversed (-1), masked (0) or retained $(+1)$. The insight behind such perturbation is that even though part of semantic information is masked out or reversed, the preserved semantic should be informative enough to coarsely highlight the change. Besides, to prevent too many channels from being masked, we define a mask ratio $\tau \in [0, 1]$ to control the ratio of 0 values in $\mathbf{v}$. It should be noted that channel-wise modulation is only adopted during training and removed from inference.

The modulated difference $\hat{\mathcal{O}}^l$ is fed into an ASPP [Chen *et al.*, 2018]-like structure to obtain a one-dimensional coarse mask $\mathcal{M}^l \in \mathbb{R}^{H_l \times W_l}$

$$\mathcal{M}^l = \sigma\left(\mathbf{MLP}\left(\mathbf{Cat}\left(\mathbf{GAP}(\hat{\mathcal{O}}^l), \hat{\mathcal{O}}^l\right)\right)\right), \tag{7}$$

where $\mathbf{GAP}(\cdot)$ denotes global average pooling and $\mathbf{Cat}(\cdot,\cdot)$ indicates concatenation between two features along channel dimension, note that the pooled feature is broadcasted to all pixels to align the resolution, $\mathbf{MLP}(\cdot)$ represents multilayer perceptron mapping the input to single channel mask, and a sigmoid function $\sigma(\cdot)$ is adopted to obtain coarse-level change area in l-th stage. In the end, GIDE feeds back such difference-dominant information to the original siamese-aligned feature

$$\widetilde{\mathcal{F}}_0^l = \hat{\mathcal{F}}_0^l \otimes \mathcal{M}^l \quad \widetilde{\mathcal{F}}_1^l = \hat{\mathcal{F}}_1^l \otimes \mathcal{M}^l \quad \mathcal{O}^l = (\hat{\mathcal{F}}_0^l - \hat{\mathcal{F}}_1^l) \otimes \mathcal{M}^l, \tag{8}$$

where $\otimes$ indicates spatial-wise modulation, i.e. $\mathcal{M}^l$ is broadcast to different channels.

Besides, we inject additional supervised objectives in the perturbation module as deep supervision to ensure the accuracy of estimated guidance mask. Concretely, we downsample the original change map $\mathcal{Y}$ to adapt the size of mask $\mathcal{M}^l$ as $\mathcal{Y}^l$ and apply the binary cross-entropy as loss

$$\begin{aligned} \mathcal{L}_d^l = &-\frac{1}{H_l W_l}\sum_{(i,j)} \mathcal{Y}^l[i,j] log(\mathcal{M}^l[i,j]) \\ &-\frac{1}{H_l W_l}\sum_{(i,j)} (1-\mathcal{Y}^l[i,j]) log(1-\mathcal{M}^l[i,j]). \end{aligned} \tag{9}$$

The deep supervision together with our designed random perturbation helps train more generalized modules to differentiate various change patterns.

3.3 Decouple Decoders

In the decoder part, to exploit the complementary property of image content and pure difference, we devise an asymmetric dual-decoder structure, which takes $\mathcal{F}_0^N, \mathcal{F}_1^N$ and $\mathcal{O}^l$ as input and predicts the change areas.

The content-aware decoder takes the encoded image features as input, and hierarchically appends difference information to generate the intermediate feature, finally a segmentation head is applied on the final output to obtain the prediction

$$\begin{aligned} \mathcal{D}_{aw}^N &= \mathbf{MLP}\left(\mathbf{Cat}(\mathcal{F}_0^N, \mathcal{F}_1^N)\right) \\ \mathcal{D}_{aw}^{l-1} &= \mathbf{Dec}_{l,aw}\left(\mathcal{S}_{\uparrow}(\mathcal{D}_{aw}^l), \mathcal{O}^{l-1}\right) \\ \hat{\mathcal{Y}}_{aw} &= \mathcal{T}_{aw}(\mathcal{D}_{aw}^1), \end{aligned} \tag{10}$$

where $\mathbf{Dec}_{l,aw}(\cdot,\cdot)$ is the l-th decoding block of decoder, which consists of concatenation and two cascaded Conv-BN-ReLU blocks with kernel size 3×3, $\mathcal{S}_{\uparrow}(\cdot)$ represents the upsampling operation, and $\mathcal{T}_{aw}(\cdot)$ is the segmentation head. As for the content-agnostic decoder, we only feed the pure difference information as input, the other structure is similar to the content-aware one

$$\begin{aligned} \mathcal{D}_{ag}^{N} &= \mathbf{MLP}\left(\mathcal{O}^{N}\right) \\ \mathcal{D}_{ag}^{l-1} &= \mathbf{Dec}_{l,ag}\left(\mathcal{S}_{\uparrow}(\mathcal{D}_{ag}^{l}), \mathcal{O}^{l-1}\right) \\ \hat{\mathcal{Y}}_{ag} &= \mathcal{T}_{ag}(\mathcal{D}_{ag}^{1}), \end{aligned} \tag{11}$$

where the structure of $\mathbf{Dec}_{l,ag}(\cdot,\cdot)$ is similar to that of content-aware decoder but the concatenation is replaced by summation of two input features. Finally, both output are summed up with an activation function to obtain final results

$$\hat{\mathcal{Y}} = \sigma(\hat{\mathcal{Y}}_{ag} + \hat{\mathcal{Y}}_{aw}). \tag{12}$$

3.4 Loss Function

During training, we simply take the cross-entropy $\mathcal{L}_{ce}$ to supervise the final output $\hat{\mathcal{Y}}$ and the total loss function can be expressed as Equation (13) with balance factor λ_1 and λ_2

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_1 \sum_{l} \mathcal{L}_{d}^{l} + \lambda_2 \mathcal{L}_{comp}. \tag{13}$$

In Equation (13), we introduce an additional comparative loss $\mathcal{L}_{comp}$ for feature regularization, which is expressed as

$$\begin{aligned} \mathcal{L}_{comp} =& \frac{1}{HW} \sum_{(i,j)} \mathcal{Y}^{N}[i,j] \left[\left\| \widetilde{\mathcal{F}}_{0}^{N}[i,j] - \widetilde{\mathcal{F}}_{1}^{N}[i,j] \right\|_{2} - \gamma \right]_{+} \\ &+ \frac{1}{HW} \sum_{(i,j)} (1 - \mathcal{Y}^{N}[i,j]) \left\| \widetilde{\mathcal{F}}_{0}^{N}[i,j] - \widetilde{\mathcal{F}}_{1}^{N}[i,j] \right\|_{2}, \end{aligned} \tag{14}$$

where $[\cdot]_{+}$ represents clipped by 0 if the value inside is negative and γ is a hyperparameter. This comparative term helps backbone to distinguish feature of changed objects.

4 Experiment

4.1 Experiment Setup

Dataset. We evaluate our proposed APD change detector on three publicly change detection datasets: LEVIR-CD [Chen and Shi, 2020a], WHU-CD [Ji *et al.*, 2018] and DSIFN-CD [Zhang *et al.*, 2020a]. LEVIR-CD is a public large building change detection dataset that contains 637 bitemporal RS image pairs of resolution 1024×1024. We utilize the default train/val/test split. WHU-CD is another public building change detection dataset, which consists of one pair of ultra-resolution (0.075m) aerial images of size 32507×15354. We follow the default cropping policy of size 512×512 and dataset split(train/test) provided by the authors. DSIFN-CD contains the changes in six major cities' land-cover objects in China. We divide the 512×512 images into 256×256 pixel patches without overlapping, and we follow the default standard train/val/test split. Consequently, there are 14400/1360/192 samples for training/val/test.

Implementation Details. We implemented our model under the Pytorch framework, using a single NVIDIA GeForce GTX 1080 Ti GPU for training and the batch size is set to 8. During training, we apply data augmentation through random flip, crop and photometric distortion. We use AdamW with weight decay equal to 0.05 for optimization. The initial learning rate is 0.001 and we train models for 60k, 40k and 100k iterations for LEVIR-CD, WHU-CD and DSIFN-CD datasets. For context aggregation, we set $k = 1$ and $d = 16$. The hyperparameters in loss term are $\lambda_1 = 1.0, \lambda_2 = 1.0, \gamma = 1.0$.

Evaluation Metrics. For horizontal comparison with other methods, we follow the common setting and use the F1 score and Intersection over Union (IoU) with regard to the change objects as the primary evaluation indices. Meanwhile, we also report precision(P) and recall(R) of the change objects.

4.2 Main Result

Methods for Comparison. To verify the effectiveness of our method, we make comparison with several advanced change detection approaches, including three purely convolutional-based methods (FC-EF [Daudt *et al.*, 2018a], FC-Siam-Conc [Daudt *et al.*, 2018a], FC-Siam-Diff [Daudt *et al.*, 2018a]), three attention-aided methods (DTCD-SCN [Liu *et al.*, 2019b], STANet [Chen and Shi, 2020b], SNUNet [Fang *et al.*, 2021]) and three methods with transformer-like structure (BIT [Chen *et al.*, 2022b], ChangeFormer [Bandara and Patel, 2022], TransUNetCD [Li *et al.*, 2022]).

Quantitative Results. Table 1 reports the overall quantitative comparison results on LEVIR-CD, DSIFN-CD and WHU-CD. In the datasets of DSIFN-CD and WHU-CD, our proposed APD change detector outperforms other methods, reaching the best level in terms of all four metrics. In the LEVIR-CD dataset, although our method does not achieve the best precision, APD can still detect more changing pixels, which makes significant advantages in the other three metrics. For example, the F1 score of our method exceeds the latest ChangeFormer by 1.31%, 1.2%, and 9.57% on the three datasets, respectively. In Table 1, we also indicate the information of the utilized backbone of different CD methods. It can be observed that APD only applies a simple and lightweight ResNet-18 network as a feature extractor and does not use complex structures such as UNet[Daudt *et al.*, 2018a; Fang *et al.*, 2021] or transformer-based network[Bandara and Patel, 2022], which are widely used in segmentation tasks. On the other hand, although our approach only takes ResNet18 as the backbone, it can still outperform competitors with larger model capacity (ResNet50) or more advanced structure (MiT-B1), which indirectly manifests the superiority of our proposed GIDE module and decoupled decoders.

Qualitative Results. In addition, Figure 3 also shows the visualization comparison of the different change detection methods on the three datasets. As highlighted in red, blue and yellow respectively, our proposed method captures more detailed change compared with other change detection schemes. In the visualization results of LEVIR-CD dataset, our APD detector can not only detect the building change more accurately, but also avoid some noise (e.g. changes in the ap-

Method	Backbone	LEVIR-CD				DSIFN-CD				WHU-CD			
		Precision	Recall	F1	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1	IoU
FC-EF	UNet	86.91	80.17	83.40	71.53	72.61	52.73	61.09	43.98	71.63	67.25	69.37	53.11
FC-Siam-Diff	UNet	89.53	83.31	86.31	75.92	59.67	65.71	62.54	45.50	47.33	77.66	58.81	41.66
FC-Siam-Conc	UNet	91.99	76.77	83.69	71.96	66.45	54.21	59.71	42.56	60.88	73.58	66.63	49.95
SNUNet	UNet++	89.18	87.17	88.16	78.83	60.60	72.89	66.18	49.45	85.60	81.49	83.50	71.67
DTCDSCN	SE-Res34	88.53	86.83	87.67	78.05	53.87	77.99	63.72	46.76	63.92	82.30	71.95	56.19
STANet	ResNet18	83.81	**91.00**	87.26	77.40	67.71	61.68	64.56	47.66	79.37	85.50	82.32	69.95
BiT	ResNet18	89.24	89.37	89.31	80.68	68.36	70.18	69.26	52.97	86.64	81.48	83.98	72.39
TransUNetCD	ResNet50	92.43	89.82	91.11	83.67	71.55	69.42	66.62	57.95	93.59	89.60	93.59	84.42
ChangeFormer	MiT-B1	92.05	88.80	90.40	82.48	88.48	84.94	86.67	76.48	89.12	82.73	85.61	75.14
Ours	ResNet18	**92.81**	90.64	**91.71**	**84.69**	**89.39**	**86.40**	**87.87**	**78.36**	**95.10**	**95.26**	**95.18**	**90.80**

Table 1: Quantitative comparison results of different change detection methods on LEVIR-CD, DSIFN-CD and WHU-CD

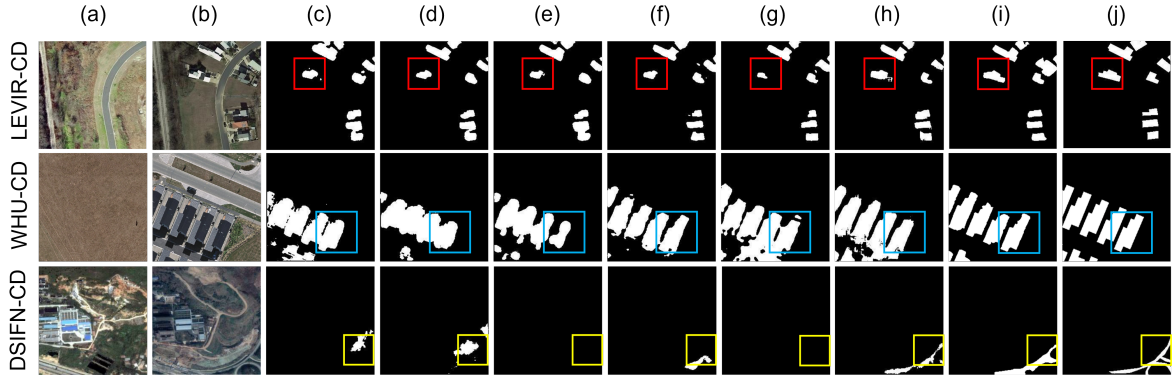

Figure 3: Qualitative results of different CD methods on LEVIR-CD, DSIFN-CD and WHU-CD dataset: (a) Pre-change image, (b) Post-change image, (c) FC-EF, (d) FC-Siam-Diff, (e) FC-Siam-Conc, (f) DTCDSCN, (g) BIT, (h) ChangeFormer, (i) Ours, and (j) Ground-truth.

Align	Perturb	Decouple	LEVIR-CD			
			Precision	Recall	F1	IoU
			85.59	86.27	85.93	75.33
✓			87.06	87.16	87.11	77.16
	✓		87.92	87.43	87.68	78.06
		✓	**93.83**	87.31	90.45	82.56
✓	✓		92.07	90.02	91.03	83.54
✓		✓	93.00	89.50	91.22	83.36
	✓	✓	93.38	89.27	91.28	83.95
✓	✓	✓	92.81	**90.64**	**91.71**	**84.69**

Table 2: Ablation study on the effectiveness of operations proposed in our approach.

pearance of land cover, seasonal illumination changes, etc.) that affects the contour of the changed area. From the visualization results of the WHU-CD dataset, it can be seen that most compared methods cannot eliminate the pseudo-change caused by shadows when detecting changes. In contrast, our method can eliminate such pseudo-change, which proves that our method can learn effective context, eliminate the irrelevant change and express the real semantic variation. At the same time, our method can also effectively detect subtle changes, which are observed from the visualization results of the DSIFN-CD dataset. The compared methods can hardly detect the detailed changed area of the long and narrow road due to the lack of precise semantic difference information. However, our APD detector can effectively capture the subtle variation and generate a more precise change map, which verifies the superiority of APD detector.

4.3 Ablation Study

Verification Experiment on Proposed Operations. We design ablation experiments for each proposed operation, i.e. Alignment, Perturbation and Decouple. For each operation, we provide a specific baseline counterpart to demonstrate the effectiveness. The overall results are shown in Table 2. The experiment shows that the Alignment module, Perturbation-aided difference module and Decoupled Decoder are helpful for change detection performance.

To evaluate the alignment operation, we remove the context-aided alignment in GIDE on purpose and directly fed the siamese features into difference prediction. In terms of Table 2, the context-aided alignment significantly improves the recall and IoU score. This is because the Alignment module effectively utilizes the contextual similarity to gather information from areas of the same semantic to enhance the semantic difference information. This improvement is also reflected in Figure 4, if the alignment part in the GIDE module is removed, the noise caused by the lack of the contextual information makes the model unable to distinguish between real changes and pseudo-changes, thus affecting the precise recognition of the changed areas.

To verify the effectiveness of perturbation, we also devise a counterpart by eliminating the deep supervision and semantic perturbation, instead we directly take the feature from both images and their difference as output. Table 2 demonstrates that our proposed perturbation-aided deep supervision enssentially improve the recall and F1-score, since combina-

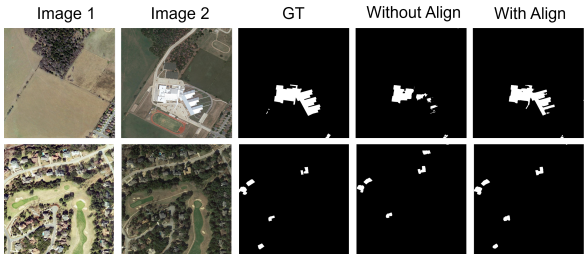

Figure 4: Visualization results of effect from Alignment

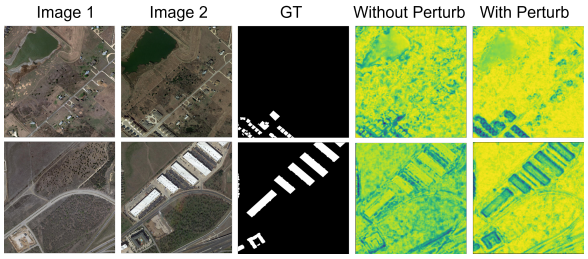

Figure 5: Feature visualization results of effect from Perturbation

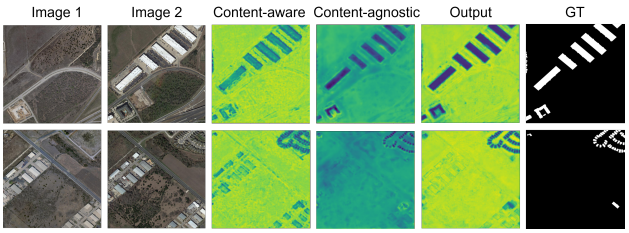

Figure 6: Feature visualization of effect from Decoupled Decoder

Deep Supervision	Mask Ratio	LEVIR-CD Precision	Recall	F1	IoU
DiceLoss	1/2	86.78	86.89	86.78	76.74
	1/4	87.65	87.23	87.44	77.68
	1/8	86.64	87.37	87.00	76.99
	1/16	86.80	87.41	87.10	77.15
CrossEntropyLoss	1/2	87.51	87.5	87.5	77.78
	1/4	**87.92**	87.43	**87.68**	**78.06**
	1/8	87.60	**87.5**	87.55	77.86
	1/16	86.90	87.35	87.12	77.18

Table 3: Ablation study on deep supervision and perturbation ratios used in the perturbation operation

tion between deep supervision and random disturbance can help difference predictor focus on more general patterns of changeful objects, consequently the estimated change area provides in-depth guidance for following feature extraction. In additional, we also visualize the heatmap of intermediate feature to show the impact from perturbation. Figure 5 shows the feature visualization before and after introducing the perturbation in the second stage of the backbone. It can be observed that with the perturbation-aided module, our model significantly reduces the impact of the pseudo-change (pond, road, etc.) and strengthens focus on the changeful objects.

Finally, we evaluate the impact from decoupled decoder, specifically, to build the baseline, we introduce a single decoder structure which takes the concatenation of image features and difference information as input. From the Table 2, we find decoupled structure brings substantial performance gain on detection precision and IoU score. This demonstrates that decoupled decoders make full use of the complementary between image content and difference information for better change detection. Figure 6 shows the feature visualization of two shunts in our decoupled decoders. We observe that the output of the Content-agnostic Decoder can retains the profile of real semantic change. Further, the dual-decoder structure can help detect some small and detailed changes that could not be detected by pure difference information.

Verification Experiment on Hyperparameters. Next we also conduct parameter verification experiments to test sensitivity to hyperparameters in the disturbance module on the LEVIR-CD dataset. The results are shown in Table 3.

First we evaluate the effect of the ratio of the masked channels, concretely, we set the ratio τ to 50%, 25%, 12.5% and 6.25% of the input feature dimension. The experimental results show that there exist negative impact on detection performance when the perturbation ratio is too high or too low, this is because large perturbation ratio can result in too many masked channels and loss in difference information, thus the difference prediction underfit to underlying change patterns and affect feature extraction in following stages. On the other hand, too few perturbed channels will decrease the variance in difference information, thus the difference prediction overfit to specific change patterns. Therefore, we choose 25% as the most appropriate ratio.

Secondly, we evaluate the effect of different forms of deep supervision $\mathcal{L}_d^l$. From Table 3, when replacing cross-entropy with dice loss, the results are degraded in different ratios of perturbation, because small objects account for a certain proportion in the change detection dataset, to which the dice loss is very sensitive. Therefore, once some pixels of small objects are incorrectly predicted, it will lead to large fluctuation in the backward gradient. Instead, the cross-entropy takes each pixel equally, which alleviates this problem.

5 Conclusion

In this paper, we review the issue of underutilization of difference information in previous methods of RSI change detection, and propose a new change detector termed as APD. The APD change detector features with three elaborately designed operations, i.e. Alignment, Perturbation and Decoupling, to fully leverage the difference information and bootstrap change detection results. With a lightweight backbone, our APD detector can effectively improve the performance on challenging benchmarks of change detection and achieve state-of-the-art results in most metrics, and extensive ablation studies also verify the effectiveness of each operation.

Acknowledgements

This work was supported in part by Natural Science Foundation of China under contract 62171139, and in part by Zhongshan science and technology development project under contract 2020AG016.

References

[Bandara and Patel, 2022] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. *arXiv preprint arXiv:2201.01293*, 2022.

[Chen and Shi, 2020a] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020.

[Chen and Shi, 2020b] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote. Sens.*, 12:1662, 2020.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[Chen *et al.*, 2022a] Hao Chen, Wenyuan Li, Song Chen, and Zhenwei Shi. Semantic-aware dense representation learning for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022.

[Chen *et al.*, 2022b] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:3095166, 2022.

[Daudt *et al.*, 2018a] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067. IEEE, 2018.

[Daudt *et al.*, 2018b] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. Ieee, 2018.

[Fang *et al.*, 2021] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hu *et al.*, 2017] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2011–2023, 2017.

[Hussain *et al.*, 2013] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of photogrammetry and remote sensing*, 80:91–106, 2013.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015.

[Ji *et al.*, 2018] Shunping Ji, Shiqing Wei, and Meng Lu. 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, 2018.

[Ji *et al.*, 2019] Shunping Ji, Yanyun Shen, Meng Lu, and Yongjun Zhang. Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples. *Remote. Sens.*, 11:1343, 2019.

[Krizhevsky *et al.*, 2017] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[Li *et al.*, 2021] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021.

[Li *et al.*, 2022] Qingyang Li, Ruofei Zhong, Xin Du, and Yuying Du. Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

[Liu *et al.*, 2019a] Ruoyun Liu, Monika Kuffer, and Claudio Persello. The temporal dynamics of slums employing a cnn-based change detection approach. *Remote. Sens.*, 11:2844, 2019.

[Liu *et al.*, 2019b] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, 18:811–815, 2019.

[Nemoto *et al.*, 2017] Keisuke Nemoto, Ryuhei Hamaguchi, M. Sato, Aito Fujita, Tomoyuki Imaizumi, and Shuhei Hikosaka. Building change detection via a combination of cnns using only rgb aerial imageries. In *Remote Sensing*, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Verma *et al.*, 2019] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

[Wang *et al.*, 2019] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[Yang *et al.*, 2022] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022.

[Zhang and Shi, 2020] Min Zhang and Wenzhong Shi. A feature difference convolutional neural network-based change detection method. *IEEE Transactions on Geoscience and Remote Sensing*, 58:7232–7246, 2020.

[Zhang *et al.*, 2020a] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. 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, 2020.

[Zhang *et al.*, 2020b] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. 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, 2020.

[Zheng *et al.*, 2021] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15193–15202, 2021.